CS 435, 2016
Lecture 1, Date: 22 February 2018
Instructor: Nisheeth Vishnoi

Preliminaries, Convexity, Duality

In this lecture, we develop the basic mathematical preliminries and tools to study convex optimization. These include some standard facts from multivariate calculus and linear algebra, convex sets and functions, and the important notion of duality.

Contents

1 Preliminaries

1.1 Calculus

We are concerned with functions $f : \mathbb{R}^n \to \mathbb{R}$. By $x, y, \ldots$ we typically mean vectors in $\mathbb{R}^n$. For two vectors x, y we use $\langle x, y \rangle$ or $x^\top y$ to denote the inner product between them and $\|x\|$ to denote the Euclidean norm. When f is smooth enough, we can talk about its gradient and Hessian. Unless otherwise stated, in this lecture we assume that the function is sufficiently smooth and not worry about its differentiability. The derivative of $f(x_1, x_2, \ldots, x_n)$ is an n-dimensional vector and is called the *gradient*. It is defined as:

$$\nabla f(x) = \left[\frac{\partial f}{\partial x_1}(x), \frac{\partial f}{\partial x_2}(x), \cdots \frac{\partial f}{\partial x_n}(x) \right]^\top.$$

All second order derivatives of the multivariate function f can be summarized in the so-called *Hessian* matrix whose element at row i and column j is

$$(\nabla^2 f(x_1, x_2, ..x_n))_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}(x).$$

In other words $\nabla^2 f(x)$ is the following $n \times n$ matrix:

$$\begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}.$$

The Hessian is symmetric as the order of i and j does not matter in differentiation.[1]

It is often useful to consider a linear or quadratic approximation of such a function around a certain point $a \in \mathbb{R}^n$, for that one can use the following Taylor expansion

Proposition 1 (Taylor expansion)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function that is infinitely differentiable. A Taylor series expansion of function f about $x = a$ is given as:*

$$f(x) = \underbrace{\underbrace{f(a) + \langle \nabla f(a), x - a \rangle}_{\textit{first order approximation}} + \frac{1}{2}(x-a)^\top \nabla^2 f(a)(x-a)}_{\textit{second order approximation}} + \underbrace{\ldots}_{\textit{Higher order terms}} \tag{1}$$

In many interesting cases one can prove that whenever x is close enough to a, then the higher order terms do not contribute much to the value of $f(x)$ and hence the second order (or even the first order) approximation give a good estimate of $f(x)$.

[1] This holds if we assume f is sufficiently differentiable.

We now give a number of basic facts from calculus that will be useful through out the course.

Proposition 2. *1. Given $f : \mathbb{R}^n \to \mathbb{R}$, twice differentiable, and $t \in [0,1]$. Let $g : \mathbb{R} \to \mathbb{R}$ be defined as*

$$g(t) = f(x + t(y - x)).$$

Then, the first two derivatives of g can be given with respect to f as

$$g'(t) = \langle \nabla f(x + t(y - x)), y - x \rangle, \tag{2}$$
$$g''(t) = \langle \nabla^2 f(x + t(y - x))(y - x), y - x \rangle. \tag{3}$$

2. Given $f : \mathbb{R}^n \to \mathbb{R}$, differentiable, and $t \in [0,1]$, we have

$$f(y) = f(x) + \int_0^1 g'(t)\, dt$$

3. Given $f : \mathbb{R}^n \to \mathbb{R}$, differentiable, and $t \in [0,1]$, we have

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle = \int_0^1 g''(t)\, dt$$

Proof. 1. The proof follows from applying partial differentiation.

2. The proof follows from (2) and the Fundamental Theorem of Calculus.

3. The proof follows from applying the Fundamental Theorem of Calculus to $g'(t)$ and $g''(t)$.

□

1.2 Linear algebra, matrices, and eigenvalues

By $\mathbb{R}^{n \times n}$ we denote the set of all square $n \times n$ matrices over reals. In many cases, the matrices we work with are symmetric, i.e., we consider $M \in \mathbb{R}^{n \times n}$ such that $M^\top = M$ (here $\top$ is the transpose operator). The identity matrix of size n is a square matrix of size $n \times n$ with ones on the main diagonal and zero everywhere else. It is denoted by I_n. If the dimension is clear from the context we drop the subscript n. An important subclass of symmetric matrices are positive semidefinite matrices, as introduced below.

Definition 3 (Positive Semidefinite Matrix)**.** *A symmetric matrix M is said to be positive semidefinite (PSD) if and only if for all $x \in \mathbb{R}^n$: $x^\top M x \geq 0$. This is denoted by:*

$$M \succeq 0.$$

M is positive definite (PD) iff $x^\top M x > 0$ holds for all non-zero $x \in \mathbb{R}^n$. This is denoted by:

$$M \succ 0.$$

Occasionally, we will make use of the following convenient notation: for two symmetric matrices M, N we write $M \preceq N$ iff $N - M \succeq 0$. It is not hard to prove that $\preceq$ defines a partial order on the set of symmetric matrices.

We now review the notions of eigenvalues and eigenvectors of square matrices.

Definition 4 (Eigenvalues and Eigenvectors)**.** *We say that $\lambda \in \mathbb{R}$ and $u \in \mathbb{R}^n$ is an eigenvalue-eigenvector pair of the matrix $A \in \mathbb{R}^{n \times n}$ if $Au = \lambda u$ and $u \neq 0$.*

Geometrically speaking, this means that eigenvectors are vectors that under the transformation A preserve the direction of the vector scaled by λ, the corresponding eigenvalue. The following figure illustrates this concept.

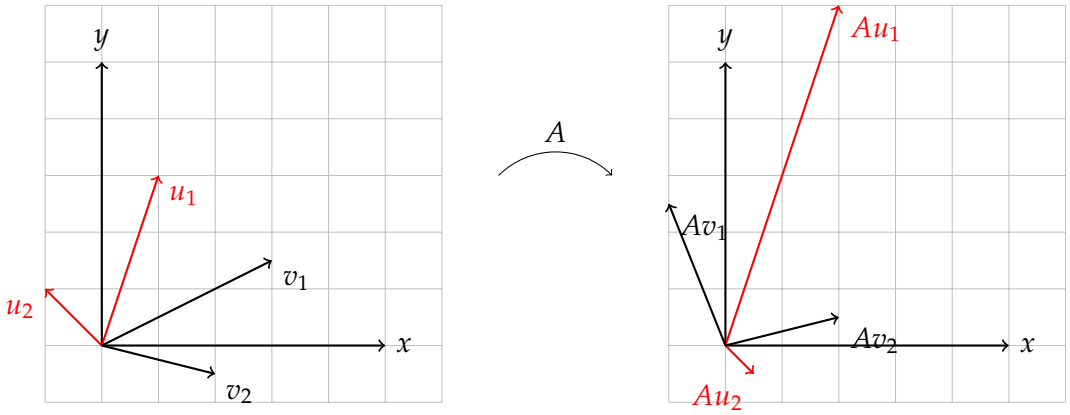


Figure 1: u_1, u_2 and $\lambda_1 = 2$, $\lambda_2 = -1/2$ are the eigenvalue-eigenvector pairs of the matrix A.

Note that for each eigenvalue λ of a matrix A with an eigenvector u, cu is also an eigenvector with eigenvalue λ.

The lemma below presents a convenient characterization of eigenvalues of a matrix.

Lemma 5. *Given $A \in \mathbb{R}^{n \times n}$ then λ is an eigenvalue of A if and only if* $\det(A - \lambda I) = 0$.

Proof. If λ is an eigenvalue then there exists some vector u such that $Au = \lambda u$. Observing that $u = Iu$ then we conclude that $(A - \lambda I)u = 0$ where 0 is the n dimensional all zero vector. Since we can assume that $u \neq 0$ then necessarily the kernel of $A - \lambda I$ is non-empty implying that its determinant is 0.

On the other hand, if $\det(A - \lambda I) = 0$ then $A - \lambda I$ is not an invertible matrix, hence there exists two vectors $u_1 \neq u_2$ such that $(A - \lambda I)u_1 = (A - \lambda I)u_2$. Re-arranging terms we find: $A(u_1 - u_2) = \lambda(u_1 - u_2)$. Therefore λ is an eigenvalue of A. □

Definition 6. *The set of eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$ of a matrix $A \in \mathbb{R}^{n \times n}$ is referred to as the spectrum of A.*

Now we will explore the properties of eigenvalues and eigenvectors of symmetric matrices.

Lemma 7. *If A is an $n \times n$ real symmetric matrix then all of its eigenvalues are real.*

Proof. If $\lambda = x + iy$ then we denote the complex conjugate of λ by $\bar{\lambda} = x - iy$, if $A \in \mathbb{C}^{m \times n}$ then we denote its complex conjugate transpose the matrix by $A^\dagger$, i.e., $A^\dagger_{ij} = \bar{A}_{ji}$. Note that if A is real valued then $A^\dagger = A^\top$.

Let λ, u be an eigenvalue-eigenvector pair of a real-valued matrix A. Then

$$u^\dagger A u = \lambda u^\dagger u = \lambda \|u\|^2 .$$

Moreover, notice that $(Au)^\dagger = u^\dagger A^\dagger = u^\dagger A^\top = u^\dagger A$. However $(Au)^\dagger = (\lambda u)^\dagger = \bar{\lambda} u^\dagger$. Therefore

$$u^\dagger A u = (uA)^\dagger u = \bar{\lambda} \|u\|^2 .$$

Since $u \neq 0$ then $\lambda = \bar{\lambda}$ which is only possible if λ is a real number. □

Lemma 8. *Given a symmetric matrix A, if λ_1, u_1 and λ_2, u_2 are two eigenvalue-eigenvector pairs such that $\lambda_1 \neq \lambda_2$ then $\langle u_1, u_2 \rangle = 0$.*

Proof. We know that $Au_1 = \lambda_1 u_1$ and $Au_2 = \lambda_2 u_2$. Moreover by symmetry we can conclude that

$$\lambda_1 u_1^\top = (Au_1)^\top = u_1^\top A^\top = u_1^\top A.$$

We also have

$$\lambda_2 u_1^\top u_2 = u_1^\top A u_2 = (Au_1)^\top u_2 = \lambda_1 u_1^\top u_2.$$

From which we conclude since $\lambda_1 \neq \lambda_2$ that $u_1^\top u_2 = \langle u_1, u_2 \rangle = 0$. □

1.3 A Useful Inequality

Vectors and the Cauchy-Schwarz Inequality. The inner product of two n-dimensional vectors $x, y \in \mathbb{R}^n$ is denoted by

$$\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i = x^\top y.$$

Unless stated otherwise, $\|x\|$ is the Euclidean norm of x. Euclidean norm is also known as ℓ_2-*norm*.

$$\|x\| = \sqrt{\langle x, x \rangle}.$$

Proposition 9 (Cauchy-Schwarz Inequality)**.** $\forall x, y \in \mathbb{R}^n$

$$\langle x, y \rangle \leq \|x\| \, \|y\|. \tag{4}$$

For vectors in $\mathbb{R}^n$, this inequality intuitively makes sense. Indeed, the two vectors x and y form together a subspace of dimension at most 2 that can be thought of as $\mathbb{R}^2$. Furtermore, assuming $x, y \in \mathbb{R}^2$, we know that $\langle x, y \rangle = \|x\| \, \|y\| \cos\theta$. Since $\cos\theta \leq 1$, the inequality holds. Nevertheless, let us go through a more algebraic proof.

Proof. The inequality can be equivalently written as

$$|\langle x, y \rangle|^2 \leq \|x\|^2 \, \|y\|^2.$$

Let us form the following non-negative polynomial in z

$$\sum_{i=1}^{n} (x_i z + y_i)^2 = \left(\sum_{i=1}^{n} x_i^2 \right) z^2 + 2 \left(\sum_{i=1}^{n} x_i y_i \right) z + \sum_{i=1}^{n} y_i^2 \geq 0.$$

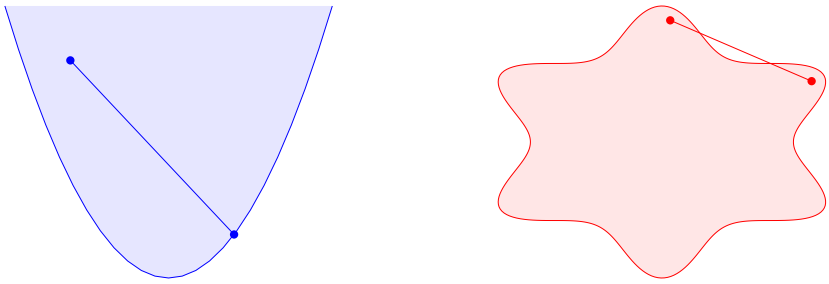
Since this degree two polynomial is non-negative it has at most one zero and its discriminant must be less than or equal to zero, implying that

$$\left(\sum_{i=1}^{n} x_i y_i \right)^2 - \sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2 \leq 0,$$

thus completing the proof. □

2 Convexity and Convex Optimization

2.1 Convex sets



(a) The shaded area is convex (b) The shaded area is not convex

Figure 2: Pictorial illustration of convexity

We start by defining the basic notion of a convex set. A set $K \subseteq \mathbb{R}^n$ is convex if $\forall x, y \in K$ and $\forall \, \lambda \in [0, 1]$ we have,

$$\lambda x + (1 - \lambda) y \in K.$$

In other words, a set $K \subseteq \mathbb{R}^n$ is convex if for every two points in K, the line segment connecting them is contained in K. For example, the set shown in Figure (b) is not convex because the line joining the two points is not fully contained in K, whsile the one in (a) is convex.

Examples.

1. Polytopes: sets of the form $K = \{x \in \mathbb{R}^n : \langle a_i, x\rangle \leq b_i \text{ for } i = 1, 2, \ldots, m\}$, where $a_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$ for $i = 1, 2, \ldots, m$.
2. Ellipsoids: sets of the form $K = \{x \in \mathbb{R}^n : x^\top Ax \leq 1\}$ where $A \in \mathbb{R}^{n\times n}$ is a positive definite matrix.
3. Balls in ℓ_p norms for $p \geq 1$, for instance $K = \{x \in \mathbb{R}^n : \sum_{i=1}^n |x_i - a_i| \leq 1\}$, where $a \in \mathbb{R}^n$ is a vector.

2.2 Convex functions

We have three ways to define a convex function. The first one holds for any function, differentiable or not.

Definition 10 (Convex function)**.** *A function $f : K \to \mathbb{R}$ is convex if $\forall x, y \in K$ and $\lambda \in [0, 1]$, we have*

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y). \tag{5}$$

Below we provide two different characterizations of convexity that might be more convenient to apply in certain situations. They hold under appropriate regularity conditions on f (differentiability or twice-differentiability respectively).

Proposition 11 (First-order convexity condition)**.** *A differentiable function $f : K \to \mathbb{R}$ over a convex set K is convex if and only if*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle, \quad \forall x, y \in K. \tag{6}$$

In other words, any tangent to a convex function f lies below the function f as illustrated in Figure 3.

Proof. We prove the one-dimensional case first. Suppose f is convex. Thus, f satisfies Definition 10 of convexity. Fix any $x, y \in K$. Then, according to 5, for every $\lambda \in [0, 1]$ we have:

$$(1 - \lambda)f(x) + \lambda f(y) \geq f(\lambda y + (1 - \lambda)x) = f(x + \lambda(y - x)).$$

Subtracting $(1 - \lambda)f(x)$ and dividing by λ yields

$$f(y) \geq f(x) + \frac{f(x + \lambda(y - x)) - f(x)}{\lambda}.$$

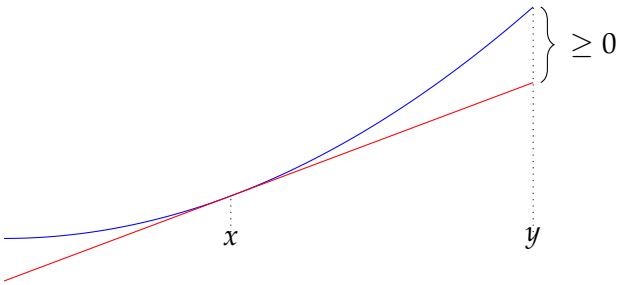


Figure 3: The first order convexity condition.

Taking limit $\lambda \to 0$, the second term on the right converges to the directional derivative in the direction $y - x$

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle.$$

Conversely, suppose the function f satisfies (6). Fix $x, y \in K$ and $\lambda \in [0,1]$. Let $z = \lambda x + (1-\lambda)y$ be some point in the convex hull. Note that the first order approximation of f around z underestimates both $f(x)$ and $f(y)$. Thus algebraically, the two underestimates are:

$$f(x) \geq f(z) + \nabla f(z)^\top (x - z), \tag{7}$$
$$f(y) \geq f(z) + \nabla f(z)^\top (y - z). \tag{8}$$

Multiplying (7) by λ and (8) by $(1-\lambda)$, and summing both inequalities, we obtain

$$\begin{aligned} (1-\lambda)f(x) + \lambda f(y)) &\geq f(z) + \nabla f(z)^\top (\lambda x + (1-\lambda)y - z) \\ &= f(\lambda y + (1-\lambda)x)). \end{aligned}$$

where the final step is to plug back the definition of z.

To extend the proof to many dimensions, just note that after fixing points $x, y \in K$ it is enough to restrict our attention to the line segment connecting them.

□

Before we go into the second order convexity conditions, we prove the lemma below that will be necessary to their proof later.

Lemma 12. *Let $f : K \to \mathbb{R}$ be a continuously differentiable function over a convex set $K \in \mathbb{R}^n$. The function f is convex iff*

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq 0. \tag{9}$$

Proof. Let f satisfy Proposition 11. Then following Proposition 2, we have

$$f(x) \geq f(y) + \nabla f(y)^\top (x - y) \text{ and } f(y) \geq f(x) + \nabla f(x)^\top (y - x).$$

Summing up both inequalities and rearranging yields (9).

Let us now assume that (9) hold for all $x, y \in K$. Let $x_\lambda = x + \lambda(y - x)$. Then

$$\begin{aligned} f(y) &= f(x) + \int_0^1 \langle \nabla f(x + \lambda(y-x)), y - x \rangle \, d\lambda \\ &= f(x) + \langle \nabla f(x), y - x \rangle + \int_0^1 \langle \nabla f(x_\lambda) - \nabla f(x), y - x \rangle \, d\lambda \\ &= f(x) + \langle \nabla f(x), y - x \rangle + \int_0^1 \frac{1}{\lambda} \langle \nabla f(x_\lambda) - \nabla f(x), x_\lambda - x \rangle \, d\lambda \\ &\geq f(x) + \langle \nabla f(x), y - x \rangle \end{aligned}$$

where the last inequality comes from the application of (9) in the integral. □

Proposition 13 (Second-order convexity condition)**.** *Suppose K is convex and open. If $f : K \to R$ is twice differentiable, then it is convex iff*

$$\nabla^2 f(x) \succeq 0, \quad \forall x \in K.$$

If the inequality is strict for all $x \in K$, the function is said to be strictly convex.

Proof. Let $f : K \to \mathbb{R}$, be twice differentiable and convex. Let $x_\tau = x + \tau s$, $\tau > 0$ and $s \in \mathbb{R}^n$. Then, from Lemma 12 we have

$$\begin{aligned} 0 \leq \frac{1}{\tau} \langle \nabla f(x_\tau) - \nabla f(x), x_\tau - x \rangle & \\ = \frac{1}{\tau} \langle \nabla f(x_\tau) - \nabla f(x), s \rangle = \frac{1}{\tau} \int_0^\tau \langle \nabla^2 f(x + \tau s) s, s \rangle \, d\lambda, & \end{aligned}$$

where the last inequality is from Proposition 2.2. The result comes from letting $\tau \to 0$.

Conversely, let $\nabla^2 f(x) \succeq 0 \; \forall x \in K$. Then

$$\begin{aligned} f(y) &= f(x) + \langle \nabla f(x), y - x \rangle \\ &\quad + \int_0^1 \int_0^\tau \underbrace{\langle \nabla^2 f(x + \lambda(y-x))(y-x), y - x \rangle}_{\geq 0} \, d\lambda d\tau \\ &\geq f(x) + \langle \nabla f(x), y - x \rangle. \end{aligned}$$

The first equality comes from Proposition 2. □

Useful convention Sometimes when working with convex functions f defined over a certain subset $K \subseteq R^n$ we will extend them to the whole $\mathbb{R}^n$ by setting $f(x) = +\infty$ for all $x \notin K$. One can check that f is then still convex (on $\mathbb{R}^n$) when the arithmetic operations on $\mathbb{R} \cup \{+\infty\}$ are interpreted in the only reasonable way.

Examples.

1. Linear functions $f(x) = \langle c, x \rangle$ for a vector $c \in \mathbb{R}^n$.
2. Quadratic functions $f(x) = x^\top Ax + b^\top x$ for a PSD matrix $A \in \mathbb{R}^n$ and a vector $b \in \mathbb{R}^n$.

2.3 Convex optimization and the usefulness of convexity

Convex optimization studies the problem of optimizing a convex function over a convex set. Given a convex set $K \subseteq \mathbb{R}^n$, and a convex function $f : \mathbb{R}^n \to \mathbb{R}$, we can define the following convex optimization problem:

$$\inf_{x \in K} f(x),^2$$

that represents the problem of finding the infimum of $f(x)$, when x is restricted to the set K. Such an optimization problem is also known as a convex program.

So far we have not yet made a case for why are we using convex functions. In this section we prove a result that illustrates why is convexity useful in optimization and give some important examples of convex programs.

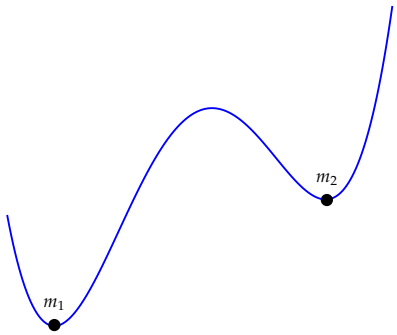


Figure 4: A function for which $\nabla f(x) = 0$ at two points m_1 and m_2.

For simplicity, let us assume an unconstrained setting i.e. $K = \mathbb{R}^n$. In Figure 4, at the points x_1 and x_2, we have $\nabla f(x_1) = 0$ and $\nabla f(x_2) = 0$, but clearly at most one of these can be the global minimum. However, if f is a convex differentiable function, we can make the following claim:

Theorem 14. *If f is a convex differentiable function, then x is an optimal solution to $\inf_{x \in \mathbb{R}^n} f(x)$ if and only if $\nabla f(x) = 0$.*

[2] We use the infimum when formalizing this optimization problem as to make sense of the case when the minimum does not exist. Typically in the problems we will be interested in the minimum does exist.

Proof. Here we prove only one direction – the one we will need. If $\nabla f(x_0) = 0$ for some x_0, then x_0 is a minimizer for f. Since f is a convex function, we know that

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \quad \forall x, y \in \mathbb{R}^n$$

Let x_0 be a point such that $\nabla f(x_0) = 0$. This implies that

$$\begin{aligned} f(y) &\geq f(x_0) + \langle \nabla f(x_0), y - x_0 \rangle \quad \forall y \in \mathbb{R}^n \\ &= f(x_0) + \langle 0, y - x_0 \rangle \quad \forall y \in \mathbb{R}^n \\ &= f(x_0). \end{aligned} \tag{10}$$

Hence, if $\nabla f(x_0) = 0$, $f(y) \geq f(x_0) \quad \forall y \in \mathbb{R}^n$, proving the claim.

□

Thus the importance of convexity is that a local minimum point is also the global minimum. Note that, in the constrained setting (i.e $K \neq \mathbb{R}^n$), Theorem 14 does not necessarily hold true. However, one can apply the following its generalization:

Theorem 15. *[1] If f is a convex differentiable function, then x is an optimal solution to $\inf_{x \in K} f(x)$ iff $\langle \nabla f(x), y - x \rangle \geq 0 \quad \forall y \in K$.*

Several important problems can be represented as convex programs. Here are two examples.

Solving Systems of Linear Equations. Suppose we are given a system of equations $Ax = b$ where $A \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^n$ with $A \succeq 0$. The traditional method to solve a system of linear equations, is through Gaussian elimination. However, the problem to solve a set of linear equations can be formulated as the following convex program.

$$\inf_x \ \frac{1}{2} x^\top A x - b^\top x.$$

As we will see in a bit, by the first order condition of optimality, we obtain

$$\nabla f(x) = Ax - b = 0,$$

implying that the optimal point to this program is the solution to the given system of equations. Is the function convex? $\nabla^2 f(x) = A$, which we is a positive definite matrix. Hence, by Proposition 13 $f(x)$ is convex. Hence solving such a convex program can lead to the solution of a system of equations.

Linear Programming. Linear programming is a technique for optimizing a linear objective function, subject to linear equality and inequality constraints. A variety of important problems can be represented as linear programs, for e.g, finding the shortest path between two vertices in a graph, or finding the maximum flow in a graph. Formally, one can represent a linear program as:
Given $A \in \mathbb{R}^{n \times m}, b \in \mathbb{R}^n, c \in \mathbb{R}^m$:

$$\begin{aligned} \inf_x \quad & \langle c, x \rangle \\ \text{subject to} \quad & Ax = b \\ & x \geq 0. \end{aligned} \tag{11}$$

The objective function $\langle c, x \rangle$ is a linear function, which is easy to prove as convex. The set of points x satisfying $Ax = b, x \geq 0$, is a polyhedron, which is a convex set. Hence, linear programs are special kinds of convex programs.

We now illustrate how the problem of finding the shortest path in a graph between two nodes can be represented as a linear program.
Given a directed graph $G = (V, E)$ with source node s, target node t, and cost w_{ij} for each edge $(i, j) \in E$, consider the program with variables x_{ij}.

$$\begin{aligned} \inf_x \quad & \sum w_{ij} x_{ij} \\ \text{subject to} \quad & x \geq 0 \\ & \forall i \quad \sum_j x_{ij} - \sum_j x_{ji} = \begin{cases} 1, & \text{if } i = s; \\ -1, & \text{if } i = t; \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{12}$$

The intuition behind this is that x_{ij} is an indicator variable for whether edge (i, j) is part of the shortest path: 1 when it is, and 0 if it is not. We wish to select the set of edges with minimal weight, subject to the constraint that this set forms a path from s to t (represented by the equality constraint: for all vertices except s and t the number of incoming and outcoming edges that are part of the path must be equal).

The equality condition can be rewritten in the form $Bx = b$, where B is the edge incidence matrix (If $e = (i, j)$ is the k^{th} edge, then the k^{th} column of B has value 1 in the j^{th} row, -1 in the i^{th} row, and 0s elsewhere) and $b \in \mathbb{R}^n$ has value 1 in the s^{th} row, -1 in the t^{th} row, and 0s elsewhere.

Note that the solution to the above linear program might be "fractional", i.e., might not give true paths. However, one can still prove that the optimal value is exactly the length of the shortest path, and one can recover a shortest path from an optimal solution to the above linear program.

3 Duality

Consider an optimization problem of the form

$$\min_{x \in K} f(x), \tag{13}$$

where, $f : \mathbb{R}^n \to \mathbb{R}$ and $K \subseteq \mathbb{R}^n$. Let $y^\star$ be its optimal value. In the process of computing $y^\star$ one often tries to obtain a good upper bound $y_U \in \mathbb{R}$ and a good lower bound $y_L \in \mathbb{R}$, so that

$$y_L \leq y^\star \leq y_U,$$

and $|y_L - y_U|$ is as small as possible. However, by inspecting the form of (13) it is evident that the problems of producing y_L and y_U seem rather different and the situation is asymmetric. Finding an upper bound y_U is much simpler, as it boils down to picking an $x \in K$ and taking $y_U = f(x)$, which is trivially a correct upper bound. Giving an (even trivial) lower bound on $y^\star$ does not seem to be such a simple task. One can think of duality as a tool to construct lower bounds for $y^\star$ in an automatic way which is almost as simple as above – it reduces to plugging in a feasible input to a different optimization problem, called the Lagrangian dual of (13).

3.1 Lagrangian Dual

Consider a problem of the form

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \ & f(x) \\ \text{s.t. } & f_j(x) \leq 0 \quad \text{for } j = 1, 2, \ldots, m. \end{aligned} \tag{14}$$

Where $f, f_1, \ldots, f_m : \mathbb{R}^n \to \mathbb{R}$ are real functions[3]. This problem is a different way to write (13) when the set K is defined by m inequalities

$$K := \{x \in \mathbb{R}^n : f_j(x) \leq 0 \ \text{ for } \ j = 1, 2, \ldots, m\}.$$

Suppose we would like to obtain a lower bound on the optimal value of (14). Towards that, one can apply the very general idea of *moving constraints to the objective*. More precisely, introduce m new variables $\lambda_1, \lambda_2, \ldots, \lambda_m \geq 0$ and consider the following Lagrangian function

$$L(x, \lambda) = f(x) + \sum_{j=1}^{m} \lambda_j f_j(x).$$

One can immediately see that, since $\lambda \geq 0$, whenever $x \in K$:

$$L(x, \lambda) \leq f(x).$$

[3]We also allow f to take the value $+\infty$, the discussion in this section is still valid in such a case.

Moreover, for every $x \in \mathbb{R}^n$, trivially

$$\max_{\lambda \geq 0} L(x, \lambda) = \begin{cases} f(x) & \text{if } x \in K \\ +\infty & \text{otherwise} \end{cases}$$

and for $x \in K$ the maximum is attained at $\lambda = 0$. Thus, consequently one can write the optimal value $y^\star$ of (14) as

$$y^\star = \min_{x \in K} \max_{\lambda \geq 0} L(x, \lambda) = \min_{x \in \mathbb{R}^n} \max_{\lambda \geq 0} L(x, \lambda).$$

It follows in particular that for every fixed $\lambda \geq 0$ we have

$$\min_{x \in \mathbb{R}^n} L(x, \lambda) \leq y^\star,$$

thus every choice of λ provides us with a *lower bound* for $y^\star$! This is exactly what we were looking for, except that now, our lower bound is a solution to an optimization problem, which is not necessarily easier than the one we started with. This is a valid concern, since we wanted a lower bound which is easy to compute. However, the optimization problem we are required to solve now is at least *unconstrained*, i.e., a the function $L(x, \lambda)$ is minimized over all $x \in \mathbb{R}^n$. In fact, for numerous important examples of problems as in (14) the value

$$g(\lambda) := \min_{x \in \mathbb{R}^n} L(x, \lambda) \tag{15}$$

has a closed-form solution (as a function of λ) and thus allows efficient computation of lower bounds to $y^\star$! We will see several examples soon.

So far we have constructed a function $g(\lambda)$ over $\lambda \geq 0$ such that for every $\lambda \geq 0$ we have $g(\lambda) \leq y^\star$. A natural question arises: what is the best lower bound we can achieve this way, i.e., what is

$$\max_{\lambda \geq 0} g(\lambda). \tag{16}$$

The above is often referred to as the *Dual Problem* (or *Dual Program*) to (14). From the above considerations we can deduce the following inequality, known as *Weak Duality*

$$\max_{\lambda \geq 0} g(\lambda) \leq \min_{x \in K} f(x). \tag{17}$$

One might ask then: does equality hold in the above inequality? It turns out that in the general case one cannot expect equality to hold, however there are sufficient conditions known, involving convexity which imply that. For instance, we have the following important theorem.

Theorem 16 (Strong Duality)**.** *Suppose that the functions $f, f_1, f_2, \ldots, f_m$ are convex and that Slater's conditions is satisfied. Then*

$$\max_{\lambda \geq 0} g(\lambda) = \min_{x \in K} f(x).$$

The Slater's condition says that there exists a point $x \in K$, such that all constraints defining K are "loose" at x, i.e., for all $j = 1, 2, \ldots, m$ we have $f_j(x) < 0$. For a proof of Theorem 16 we refer to [1]. We note that Theorem 16 is also known to hold under slightly weaker assumptions, i.e., if the set K lies on a lower-dimensional subspace, the Slater's condition should be satisfied in the *relative interior* of K only.

Example: Dual of the Linear Program in Standard Form. Consider a linear programming problem of the form

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \langle c, x \rangle \\ \text{s.t.} \quad & Ax \leq b \end{aligned} \tag{18}$$

where A is an $m \times n$ matrix and $b \in \mathbb{R}^m$ is a vector. The notation $Ax \leq b$ is a short way to say

$$\langle a_i, x \rangle \leq b_i \qquad \text{for all } i = 1, 2, \ldots, m,$$

where $a_1, a_2, \ldots, a_m$ are the rows of A.

Its dual can be seen to be

$$\begin{aligned} \max_{y \in \mathbb{R}^m} \quad & \langle b, y \rangle \\ \text{s.t.} \quad & A^\top y \geq c. \end{aligned} \tag{19}$$

3.2 Conjugate Function

A notion related to duality and often useful in deriving duals of optimization problems is the following.

Definition 17 (Conjugate function)**.** *For a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, its conjugate $f^\star : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is defined as*

$$f^\star(y) = \sup_{x \in \mathbb{R}^n} \langle y, x \rangle - f(x),$$

for $y \in \mathbb{R}^n$.

It is not hard to show (exercise) that $f^\star$ is always convex – even if f is not. Directly from the definition we can deduce the following

Proposition 18 (*Fenchel's inequality*)**.** *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be any function, then for all $x, y \in \mathbb{R}^n$ we have*

$$f(x) + f^\star(y) \geq \langle x, y \rangle.$$

Examples of Conjugate Functions

1. If $f(x) = ax + b$, then $f^\star(a) = -b$ and $f^\star(y) = \infty$ for $y \neq a$.
2. If $f(x) = \frac{1}{2}x^2$, then $f^\star(y) = \frac{1}{2}y^2$.
3. If $f(x) = x \log x$, then $f^\star(y) = e^{y-1}$.

References

[1] Lieven Vandenberghe Stephen Boyd. *Convex optimization*. Cambridge University Press, 2004.